Probabilidad y estadística

Emilio Pereyra

June 10, 2024

Este es un resumen de lo que estudio para rendir probabilidad y estadística.
Agradecimientos especiales a la Cami que me prestó la carpeta.

1 Introducción y preeliminares

1.1 ¿Que es la estadística?

Vamos a estudiar dos tipos de estadística. La estadística descriptiva y la estadística inferencial. La estadística descriptiva se encarga de resumir o presentar los datos que se toman de individuos, mientras que la inferencial se encarga de deducir información o nuevos datos estadisticos a partir de los que se tienen.

Definition 1.1 (Población) *Una población es un conjunto de individuos con alguna cualidad de interes.*

Definition 1.2 (Muestra) *Una muestra es un subconjunto de la población, que es respresentativo de esta última.*

Las muestras es sobre donde va a trabajar la estadística descriptiva, y los datos que va a presentar son sobre los individuos de una muestra.

Ahora vamos a hablar mas en concreto de los datos. En general decimos que un indivudio tiene dos tipos de datos para describirlo.

Definition 1.3 (Datos númericos) *Simplemente, un dato númerico es dado por un número. Pero diferenciamos los datos númericos en continuos y discretos.*
Ejemplos de datos númericos son: una temperatura, el tiempo, una edad, etc.

Definition 1.4 (Datos Categóricos) *Un dato categórico es un dato que describe alguna cualidad del individuo. Por ejemplo: Dificultad (Fácil, difícil,etc), color, sexo, grupo sanguíneo, etc.*

2 Estadística descriptiva

Vamos a discutir ahora justamente el como hacer estadística descriptiva. Construir histogramas, Boxplots, tablas de distribución, etc.

2.1 Tablas de frecuencia

Sup que tenemos datos $x_1, x_2, ..., x_n$.
1°) Buscamos un intervalo que contenga los datos. (Para que no se salgan del grafíco digamos)
2°) Dividir el intervalo en k intervalos de clase (IC) que sean adyacentes y disjuntos. 3°) Contar el número de observaciones de cada IC. Esto se llama frecuencia absoluta (FA).
4°) Calculamos la frecuencia relativa (FR) como $\frac{FA}{n}$. Donde n es el n° de datos.
Observar que: Para tomar el k podemos hacer: $k \approx \sqrt{n}$.
Veamos un ejemplo:
Cada entrada de las siguientes representa la cantidad de avistamientos de dodos en distintos parques nacionales.

$$\begin{array}{ccccc} 32 & 37 & 57 & 70 & 74 \\ 75 & 76 & 109 & 166 & 177 \\ 190 & 193 & 203 & 241 & 242 \\ 269 & 336 & 359 & 406 & 455 \\ 507 & 647 & 832 & 999 & 1258 \end{array} \tag{1}$$

El intervalo que vamos a elegir para el grafico va a ser $[0, 1500]$.
Tomamos $k = \sqrt{25} = 5$
Ahora para no complicarse agarramos intervalos de clase todos iguales pero esto no conviene siempre. La longitud de los IC va a ser: $\frac{b-a}{k} = \frac{1500}{5} = 300$.

Entonces la tabla de frecuencia bajo estos parametros queda:
TODO: COMPLETAR CON EL GRAFICO

2.2 Histograma

TODO: COMPLETAR

2.3 Medias de posición

Definition 2.1 (Media muestral o promedio muestral) *Dados los datos $x_1, x_2, ..., x_n$ llamamos media muestral y denotamos con $\overline{x}$ ó $\overline{x}_n = \sum_{i=0}^{n} \frac{x_n}{n}$. Esta media es de las utilizadas pero notemos que es sensible a datos extremos.*

Definition 2.2 (Percentiles muestrales) *El percentil muestral i deja al i% de los datos a izquierda y el resto a la derecha. Lo que queremos decir con*

esto es que al comparar un dato con el percentil (percentil $> x$ ó percentil $< x$), el dato queda separado junto al i% de los datos ó al porcentaje restante.

Denotamos el percentil como $\widetilde{x}$. Dada la muestra $x_1, x_2, ..., x_n$, ordenada de menor a mayor, osea que $x_1 \leq x_2 \leq ... \leq x_n$

$$Percentil\ i = \begin{cases} \frac{x_{(n/i)}+x_{(n/i)+1}}{2} & Si\ n\ es\ par \\ x_{(\frac{n+1}{i})} & Si\ n\ es\ impar \end{cases} \tag{2}$$

Un caso particular del percentil muestral es la media muestra que es simplemente tomar: $i = 50$ por lo que los datos se dividen a la mitad.

Un caso particular del percentil muestral es la mediana muestral.

Definition 2.3 (Mediana muestral) *Mediana muestral = Percentil* 50

$$Percentil\ i = \begin{cases} \frac{x_{(n/2)}+x_{(n/2)+1}}{2} & Si\ n\ es\ par \\ x_{(\frac{n+1}{2})} & Si\ n\ es\ impar \end{cases} \tag{3}$$

Definition 2.4 (Primer cuatril y tercer cuatril. Q_1,Q_3) *El primer cuatril, comunmente denotado Q_1 es simplemente la mediana de la mitad mas chica de los datos. Analogamente definimos el tercer cuatril como la mediana de la mitad mas grande de los datos.*
Dada la muestra $x_1, x_2, ..., x_n$, ordenada de menor a mayor, osea que $x_1 \leq x_2 \leq ... \leq x_n$.

$$Q_3 = \begin{cases} mediana\{x_i : 1 \leq i \leq \frac{n}{2}\} & Si\ n\ es\ par. \\ mediana\{x_i : 1 \leq i \leq \frac{n+1}{2}\} & Si\ n\ es\ impar. \end{cases}$$

Definition 2.5 (Rango) *El rango de una muestra se define como la diferencia entre el maximo valor y el minimo valor, i.e:*
Dada la muestra $x_1, x_2, ..., x_n$, ordenada de menor a mayor, osea que $x_1 \leq x_2 \leq ... \leq x_n$.

$$Rango = x_n - x_1$$

Definition 2.6 (Varianza Muestral) *La varianza muestral es una medida que nos permite dimensionar la dispersión del valor de los datos de nuestra muestra.*

Siendo que da un valor muy alto decimos que hay más dispersión que si da un valor mas pequeño.

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{4}$$

Notemos que esta medida no esta en la misma unidad que los datos. Pero ahora veremos una alternativa que salva eso.

Definition 2.7 (Desvío estándar muestral) *Se define de la siguiente forma:*

$$S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{S_n^2} \tag{5}$$

Esta forma nos da una medida en las mismas unidades que nuestros datos.

Definition 2.8 (Varianza muestral corregida y desviación estandár corregida) *Primero definamoslas y luego hablamos de porque son necesarias:*

Varianza muestral corregida:

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{6}$$

Desviación estándar muestral corregida:

$$S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{S_{n-1}^2} \tag{7}$$

A primera vista no parece tener sentido, o relación algúna la defínción de desviación estándar corregida. La realidad es que es un concepto al que debemos llegar despúes. De momento lo único que podemos decir es que estamos trabajando sobre **muestras**, la cuestión cambiará al avanzar y darnos cuenta que ver solo ver la muestra introduce un **sesgo**. De nuevo ahondaremos en esto más adelante.

Definition 2.9 (Rango intercuatril)

$$RIC = Q_3 - Q_1 \tag{8}$$

Definition 2.10 (Coeficiente de variación)

$$CV = \frac{S_n}{\overline{x}} \cdot 100\% \tag{9}$$

La ventaja de esta medidad es que no depende de las unidades medidads